Bilevel Learning of the Group Lasso Structure

Jordan Frecon[1], Saverio Salzo[1], Massimiliano Pontil[1,2]

[1] CSML - Istituto Italiano di Tecnologia
[2] Dept of Computer Science - University College London

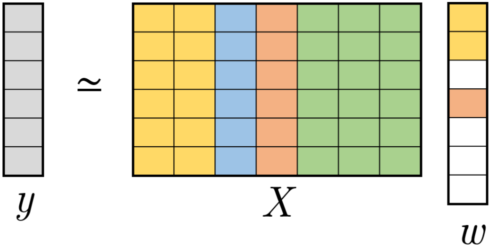
ISTITUTO ITALIANO DI TECNOLOGIA

Thirty-second Conference on Neural Information Processing Systems, Montreal, Canada

Linear Regression and Group Sparsity

Problem: Predict $y \in \mathbb{R}^N$ from $X \in \mathbb{R}^{N \times P}$

Linear Regression: Find $w \in \mathbb{R}^P$ such that



In many applications, few groups are relevant to predict $y \Rightarrow$ Group Sparse w

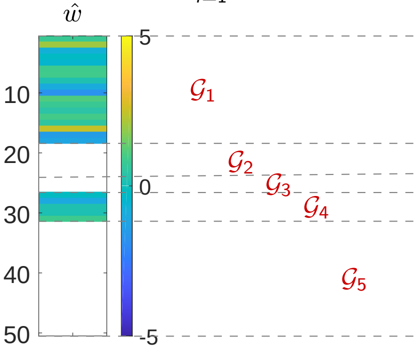
- Predict psychiatric disorder from activities in regions of the brain
- Predict protein functions from their molecular composition

Group Lasso

Given $\lambda > 0$ and a group-structure $\{\mathcal{G}_1, \ldots, \mathcal{G}_L\}$, find

$$\hat{w} \in \underset{w \in \mathbb{R}^P}{\operatorname{argmin}} \frac{1}{2}\|y - Xw\|^2 + \lambda \sum_{l=1}^{L} \|w_{\mathcal{G}_l}\|_2,$$

Group-sparse solution $\hat{w}$



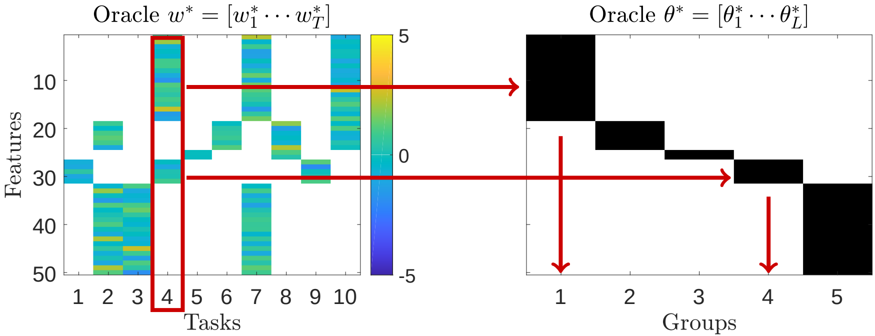
Limitation: The group-structure $\{\mathcal{G}_1, \ldots, \mathcal{G}_L\}$ may be unknown

Setting

Setting: T Group Lasso problems with shared group-structure

$$(\forall t \in \{1,\ldots,T\}) \quad \hat{w}_t(\theta) \in \underset{w_t \in \mathbb{R}^P}{\operatorname{argmin}} \frac{1}{2}\|y_t - X_t w_t\|^2 + \lambda \sum_{l=1}^{L} \|w_t \odot \theta_l\|_2,$$

encodes groups



Goal: Estimation of the optimal group-structure θ^*

A Bilevel Programming Approach

Upper-level Problem:

$$\underset{[\theta_1 \cdots \theta_L] \in \Theta}{\text{minimize}}\ \mathcal{U}(\theta) := \sum_{t=1}^{T} \mathcal{E}_t(\hat{w}_t(\theta)) \qquad \textit{(e.g., validation error)}$$

where $\hat{\mathrm{w}}(\theta) = \left[\hat{w}_1(\theta) \cdots \hat{w}_T(\theta)\right]$ solves

Lower-level Problem: (T Group Lasso problems)

$$\underset{\mathrm{w} \in \mathbb{R}^{P \times T}}{\text{minimize}}\ \mathcal{L}(w, \theta) := \sum_{t=1}^{T} \left(\frac{1}{2} \|y_t - X_t w_t\|^2 + \lambda \sum_{l=1}^{L} \|\theta_l \odot w_t\|_2 \right)$$

Difficulties:

- $\hat{\mathrm{w}}(\theta)$ not available in closed form
- $\theta \mapsto \hat{\mathrm{w}}(\theta)$ is nonsmooth [$\Rightarrow \mathcal{U}$ is nonsmooth]

Approximate Bilevel Problem

Upper-level Problem:

$$\underset{[\theta_1 \cdots \theta_L] \in \Theta}{\text{minimize}}\ \mathcal{U}_K(\theta) := \sum_{t=1}^{T} \mathcal{E}_t(w_t^{(K)}(\theta))$$

where $w_t^{(K)}(\theta) \to \hat{w}_t(\theta)$

Dual Algorithm:

$\mathrm{u}^{(0)}(\theta)$ chosen arbitrarily
for $k = 0, 1, \ldots, K-1$
$\lfloor\ \mathrm{u}^{(k+1)}(\theta) = \mathcal{A}(\mathrm{u}^{(k)}(\theta), \theta)$ *dual update*
$\left[w_1^{(K)}(\theta) \cdots w_T^{(K)}(\theta)\right] = \mathcal{B}(\mathrm{u}^{(K)}(\theta), \theta)$ *primal dual relationship*

Goals:

- Find $\mathcal{A}$ and $\mathcal{B}$ smooth [$\Rightarrow \mathrm{w}^{(K)}$ is smooth $\Rightarrow \mathcal{U}_K$ is smooth]
- Prove that the approximate bilevel scheme converges

.

Contributions

- **Bilevel Framework for Estimating the Group Lasso Structure**
- **Design of a Dual Forward-Backward Algorithm with Bregman Distances such that**

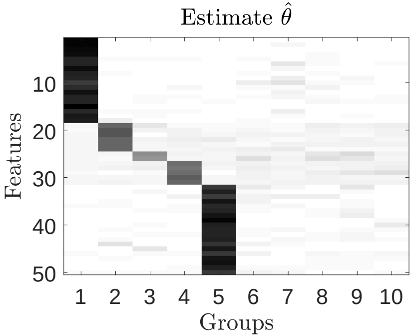
  1. $\mathcal{A}$ and $\mathcal{B}$ are smooth $\Rightarrow$ $\mathcal{U}_K$ is smooth
  2. $\begin{cases} \text{min } \mathcal{U}_K \to \text{min } \mathcal{U} \\ \operatorname{argmin} \mathcal{U}_K \to \operatorname{argmin} \mathcal{U} \end{cases}$

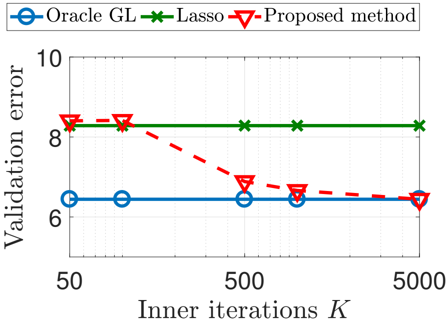
Implementation of proxSAGA algorithm: nonconvex stochastic variant of

$$\theta^{(q+1)} = \mathcal{P}_\Theta\big(\theta^{(q)} - \gamma \nabla \mathcal{U}_K(\theta^{(q)})\big)$$

Numerical Experiment

Setting: $T = 500$ tasks, $N = 25$ noisy observations, $P = 50$ features.
Estimate and group the features into, at most, $L = 10$ groups.





Conclusion

Thank You

Our poster AB #92 will be presented in Room 210 & 230 at 5pm